US009270941B1

(12) **United States Patent**
**Lavelle**

(10) **Patent No.: US 9,270,941 B1**
(45) **Date of Patent: Feb. 23, 2016**

(54) **SMART VIDEO CONFERENCING SYSTEM**

(71) Applicant: **Logitech Europe S.A**, Lausanne (CH)

(72) Inventor: **Mark Lavelle**, San Mateo, CA (US)

(73) Assignee: **LOGITECH EUROPE S.A.**, Laussane (CH)

( * ) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **14/659,373**

(22) Filed: **Mar. 16, 2015**

(51) **Int. Cl.**
***H04N 7/14*** (2006.01)
***H04N 7/15*** (2006.01)

(52) **U.S. Cl.**
CPC ................ ***H04N 7/15*** (2013.01); ***H04N 7/142*** (2013.01); ***H04N 7/147*** (2013.01)

(58) **Field of Classification Search**
CPC ................................ H04M 2203/50–2203/509
USPC .................... 348/14.01–14.16; 370/259–271, 370/351–357; 379/201.01, 202.01–207.01; 709/201–207, 217–248
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

| | | | | |
|---|---|---|---|---|
| 5,434,617 | A * | 7/1995 | Bianchi | 348/170 |
| 6,392,694 | B1 * | 5/2002 | Bianchi | 348/169 |
| 6,611,281 | B2 * | 8/2003 | Strubbe | 348/14.01 |
| 6,731,334 | B1 * | 5/2004 | Maeng et al. | 348/211.12 |
| 6,829,391 | B2 * | 12/2004 | Comaniciu et al. | 382/243 |
| 7,349,008 | B2 * | 3/2008 | Rui et al. | 348/169 |
| 7,433,327 | B2 * | 10/2008 | Harville et al. | 370/260 |
| 8,094,193 | B2 * | 1/2012 | Peterson et al. | 348/169 |
| 8,284,254 | B2 * | 10/2012 | Romanowich et al. | 348/154 |
| 8,358,328 | B2 * | 1/2013 | Friel et al. | 348/14.08 |
| 8,471,889 | B1 * | 6/2013 | Lee et al. | 348/14.07 |
| 8,780,168 | B2 | 7/2014 | Corley et al. | |
| 8,842,161 | B2 * | 9/2014 | Feng et al. | 348/14.12 |
| 8,872,882 | B2 | 10/2014 | Shanmukhadas et al. | |
| 8,885,057 | B2 | 11/2014 | Mock | |
| 8,913,103 | B1 * | 12/2014 | Sargin et al. | 348/14.12 |
| 2004/0003409 | A1 * | 1/2004 | Berstis | 725/105 |
| 2011/0128350 | A1 * | 6/2011 | Oliver et al. | 348/36 |
| 2013/0335508 | A1 * | 12/2013 | Mauchly | 348/14.08 |
| 2014/0111600 | A1 * | 4/2014 | Schaefer et al. | 348/14.08 |
| 2015/0022636 | A1 * | 1/2015 | Savransky | 348/46 |

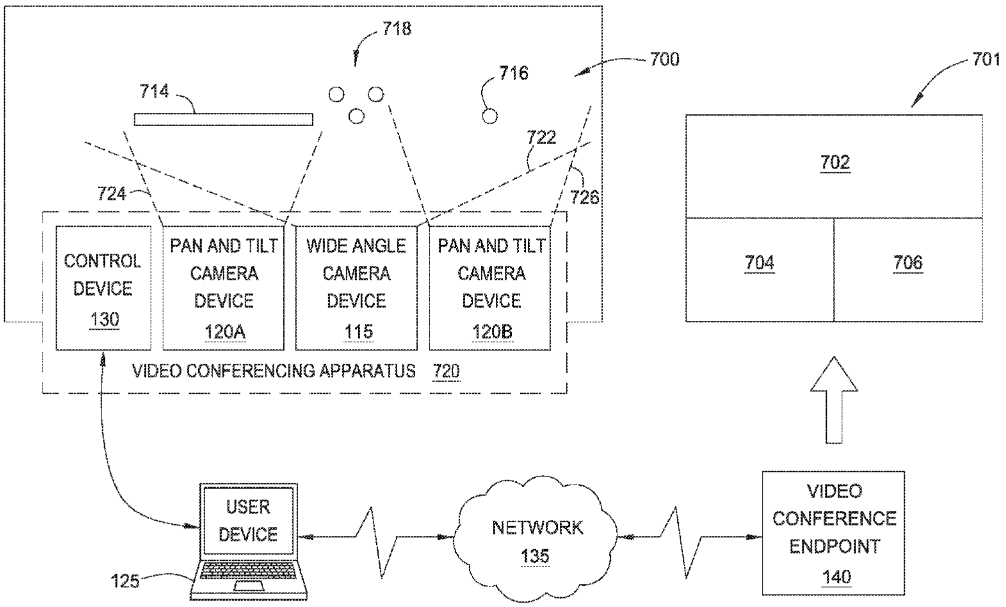
* cited by examiner

*Primary Examiner* — Hemant Patel
(74) *Attorney, Agent, or Firm* — Patterson & Sheridan, LLP

(57) **ABSTRACT**

Embodiments provide techniques for facilitating transmission of a video stream from a first video conferencing device to a remote video conferencing device. Embodiments receive, by the first video conferencing endpoint device, first video data captured from a first field of view of a physical environment. The video data includes a plurality of frames. Activity data is determined for portions of the first video data across the plurality of frames. Embodiments generate, by a first video conferencing endpoint device, second video data from a second field of view of the physical environment, based on the determined activity data. Additionally, embodiments facilitate the transmission of the video stream to the remote video conferencing device for display, the video stream comprising the generated second video data and audio data captured within the physical environment.

**29 Claims, 8 Drawing Sheets**


718
714
716
700
701
722
724
726
702
704
706
CONTROL DEVICE 130
PAN AND TILT CAMERA DEVICE 120A
WIDE ANGLE CAMERA DEVICE 115
PAN AND TILT CAMERA DEVICE 120B
VIDEO CONFERENCING APPARATUS 720
USER DEVICE
125
NETWORK 135
VIDEO CONFERENCE ENDPOINT 140

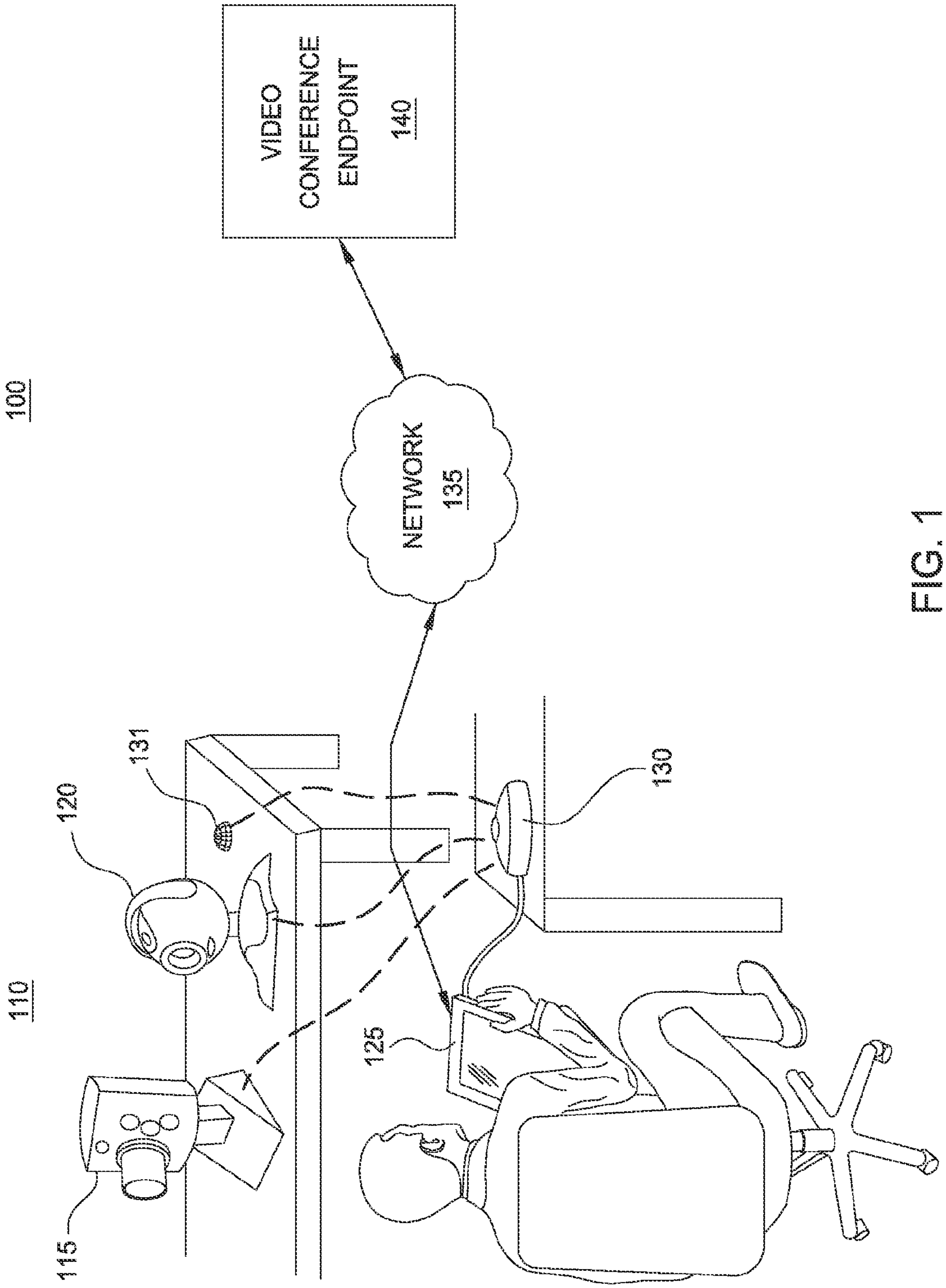

100
110
115
120
131
130
125
NETWORK
135
VIDEO
CONFERENCE
ENDPOINT
140

FIG. 1

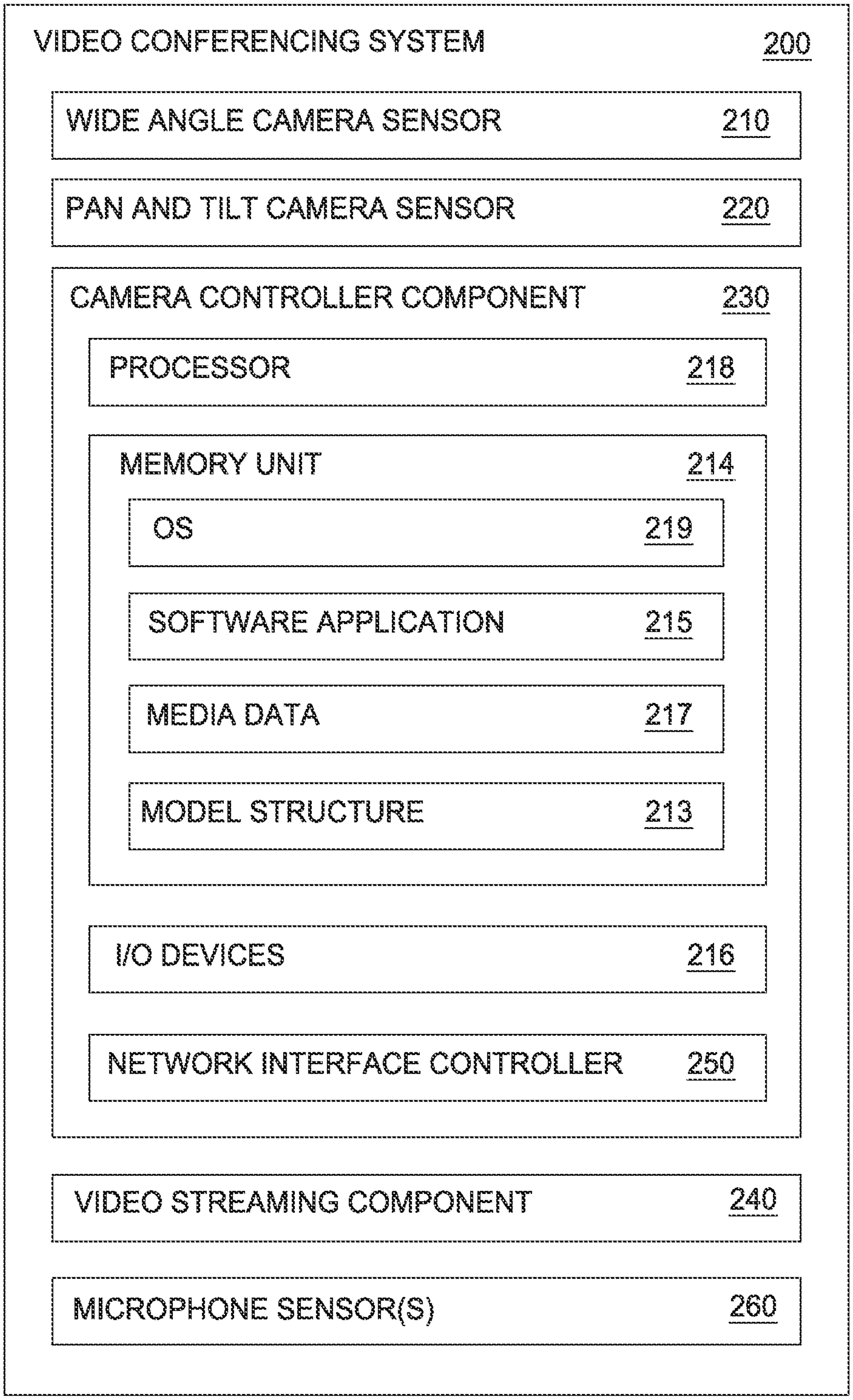

VIDEO CONFERENCING SYSTEM 200
WIDE ANGLE CAMERA SENSOR 210
PAN AND TILT CAMERA SENSOR 220
CAMERA CONTROLLER COMPONENT 230
PROCESSOR 218
MEMORY UNIT 214
OS 219
SOFTWARE APPLICATION 215
MEDIA DATA 217
MODEL STRUCTURE 213
I/O DEVICES 216
NETWORK INTERFACE CONTROLLER 250
VIDEO STREAMING COMPONENT 240
MICROPHONE SENSOR(S) 260

FIG. 2

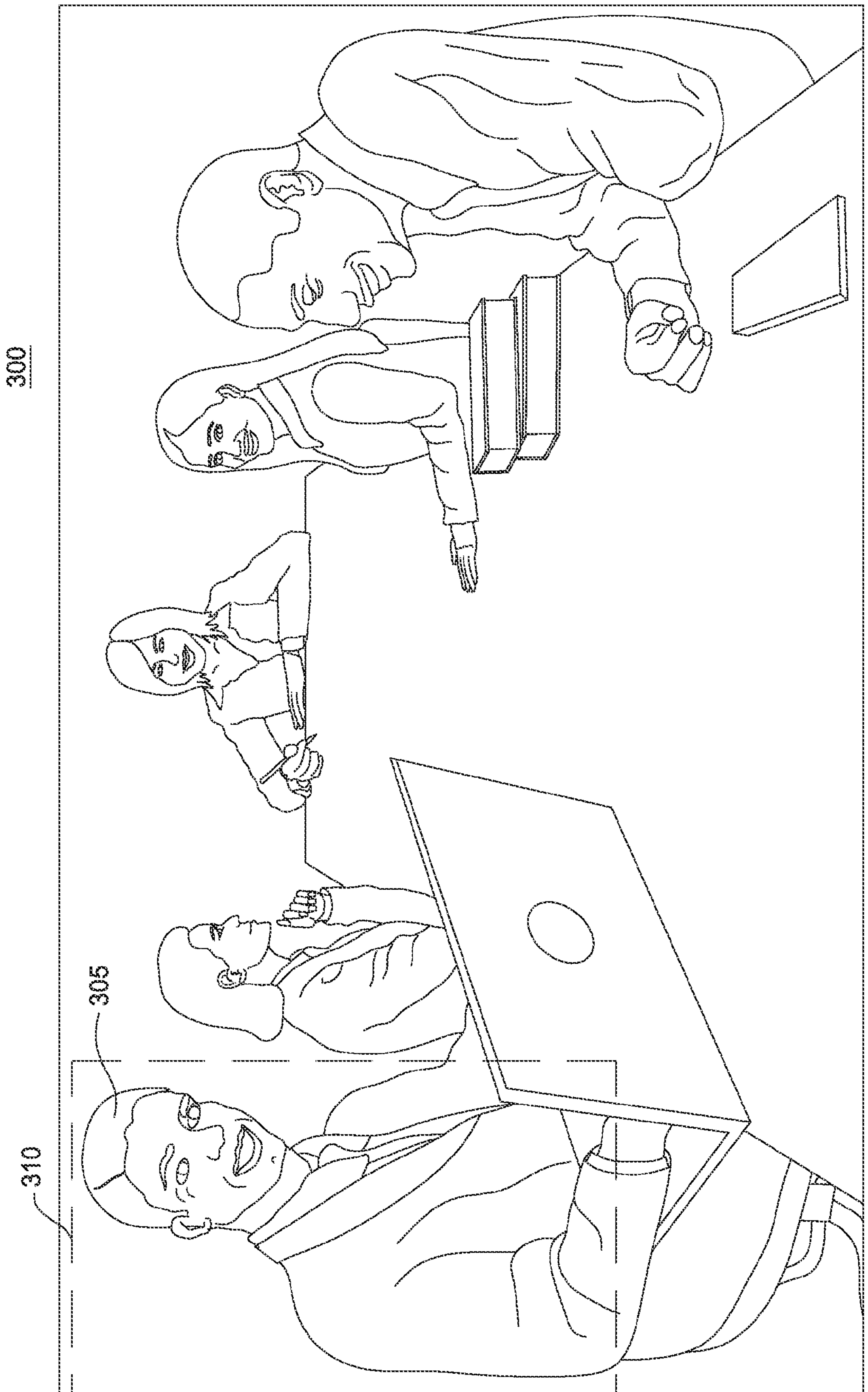


FIG. 3

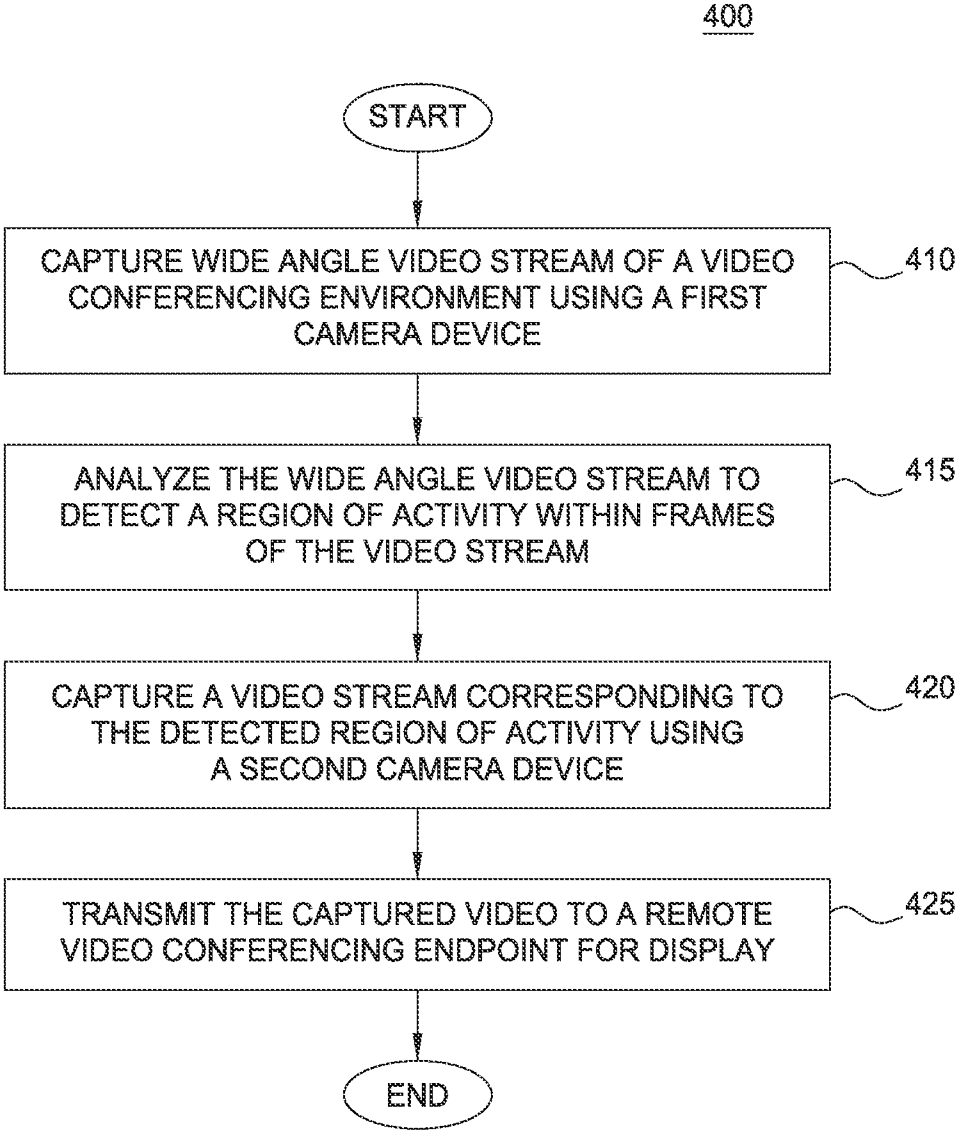

400
START
CAPTURE WIDE ANGLE VIDEO STREAM OF A VIDEO CONFERENCING ENVIRONMENT USING A FIRST CAMERA DEVICE
410
ANALYZE THE WIDE ANGLE VIDEO STREAM TO DETECT A REGION OF ACTIVITY WITHIN FRAMES OF THE VIDEO STREAM
415
CAPTURE A VIDEO STREAM CORRESPONDING TO THE DETECTED REGION OF ACTIVITY USING A SECOND CAMERA DEVICE
420
TRANSMIT THE CAPTURED VIDEO TO A REMOTE VIDEO CONFERENCING ENDPOINT FOR DISPLAY
425
END

FIG. 4

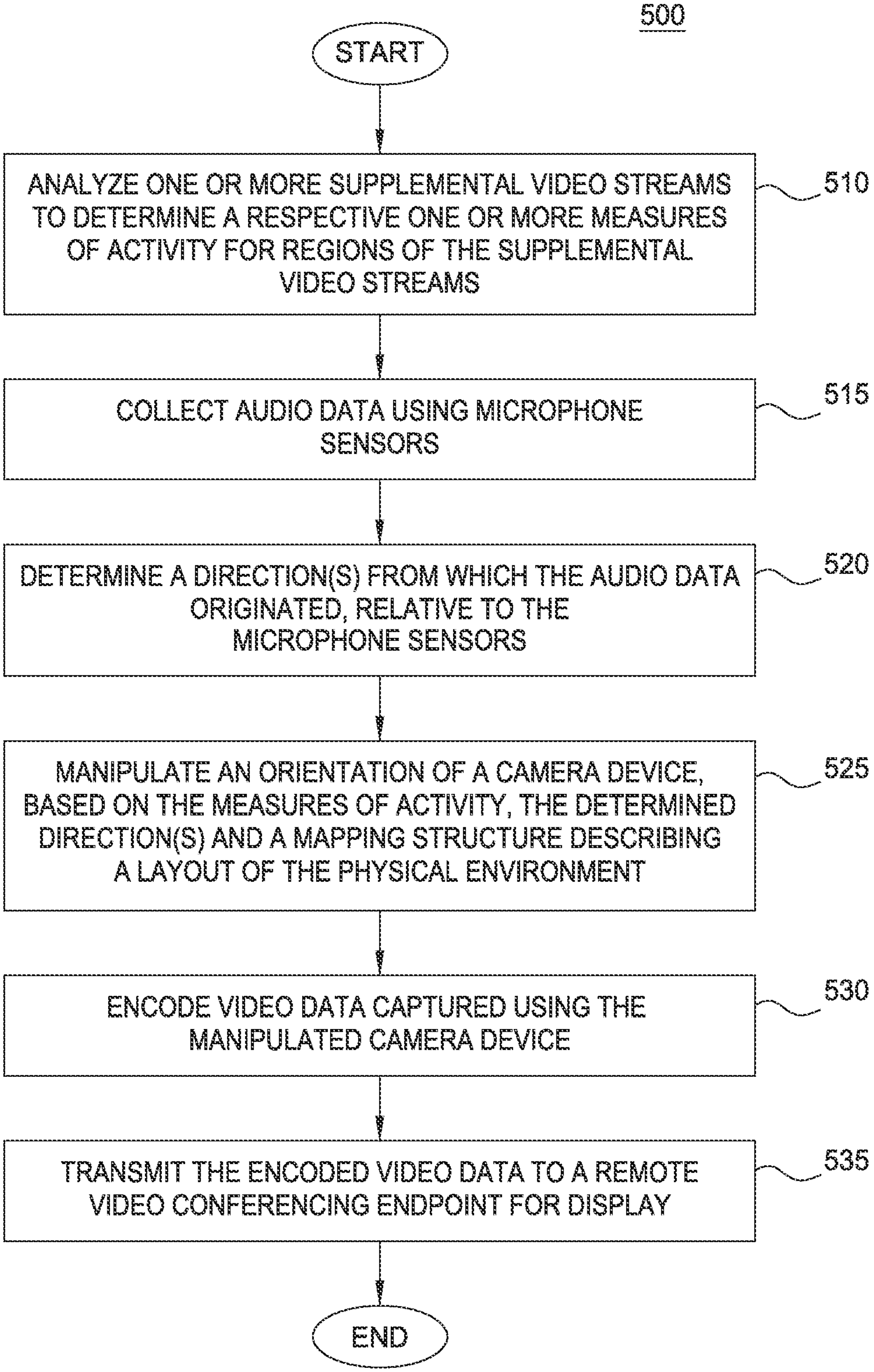

500
START
ANALYZE ONE OR MORE SUPPLEMENTAL VIDEO STREAMS TO DETERMINE A RESPECTIVE ONE OR MORE MEASURES OF ACTIVITY FOR REGIONS OF THE SUPPLEMENTAL VIDEO STREAMS
510
COLLECT AUDIO DATA USING MICROPHONE SENSORS
515
DETERMINE A DIRECTION(S) FROM WHICH THE AUDIO DATA ORIGINATED, RELATIVE TO THE MICROPHONE SENSORS
520
MANIPULATE AN ORIENTATION OF A CAMERA DEVICE, BASED ON THE MEASURES OF ACTIVITY, THE DETERMINED DIRECTION(S) AND A MAPPING STRUCTURE DESCRIBING A LAYOUT OF THE PHYSICAL ENVIRONMENT
525
ENCODE VIDEO DATA CAPTURED USING THE MANIPULATED CAMERA DEVICE
530
TRANSMIT THE ENCODED VIDEO DATA TO A REMOTE VIDEO CONFERENCING ENDPOINT FOR DISPLAY
535
END

FIG. 5

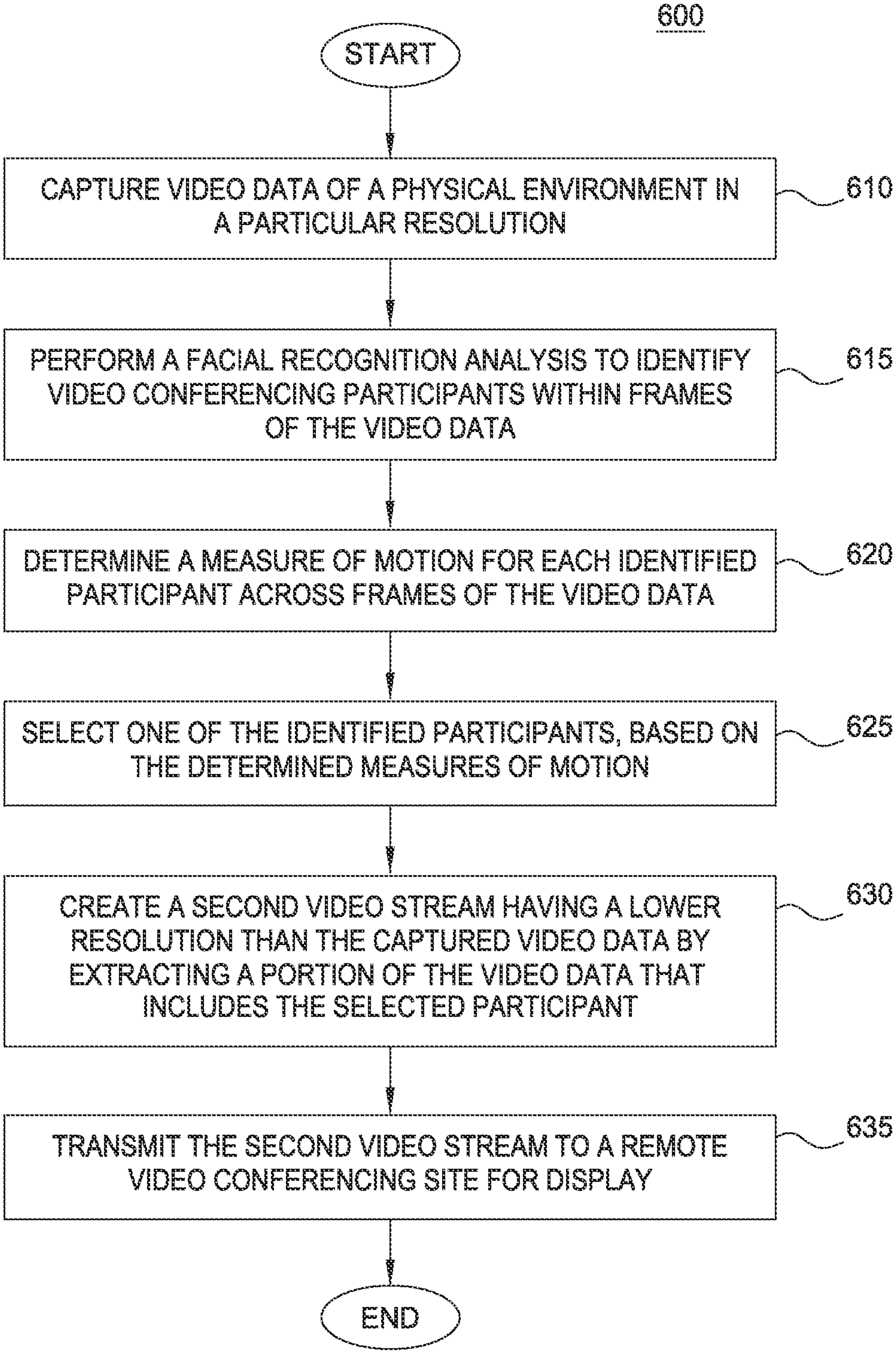

600
START
CAPTURE VIDEO DATA OF A PHYSICAL ENVIRONMENT IN A PARTICULAR RESOLUTION
610
PERFORM A FACIAL RECOGNITION ANALYSIS TO IDENTIFY VIDEO CONFERENCING PARTICIPANTS WITHIN FRAMES OF THE VIDEO DATA
615
DETERMINE A MEASURE OF MOTION FOR EACH IDENTIFIED PARTICIPANT ACROSS FRAMES OF THE VIDEO DATA
620
SELECT ONE OF THE IDENTIFIED PARTICIPANTS, BASED ON THE DETERMINED MEASURES OF MOTION
625
CREATE A SECOND VIDEO STREAM HAVING A LOWER RESOLUTION THAN THE CAPTURED VIDEO DATA BY EXTRACTING A PORTION OF THE VIDEO DATA THAT INCLUDES THE SELECTED PARTICIPANT
630
TRANSMIT THE SECOND VIDEO STREAM TO A REMOTE VIDEO CONFERENCING SITE FOR DISPLAY
635
END

FIG. 6

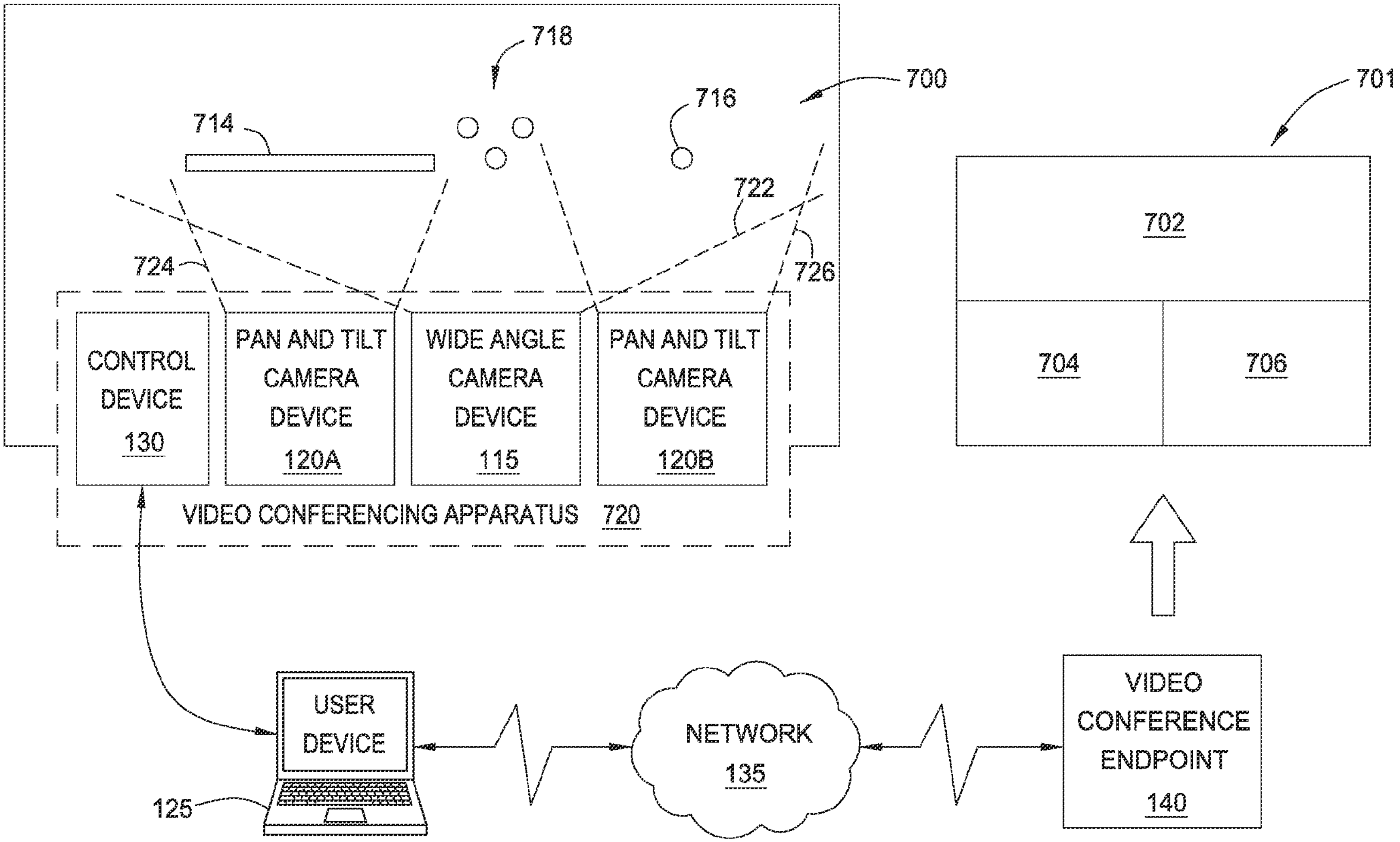

718
700
701
714
716
722
724
726
702
704
706
CONTROL DEVICE 130
PAN AND TILT CAMERA DEVICE 120A
WIDE ANGLE CAMERA DEVICE 115
PAN AND TILT CAMERA DEVICE 120B
VIDEO CONFERENCING APPARATUS 720
USER DEVICE
125
NETWORK 135
VIDEO CONFERENCE ENDPOINT 140

FIG. 7

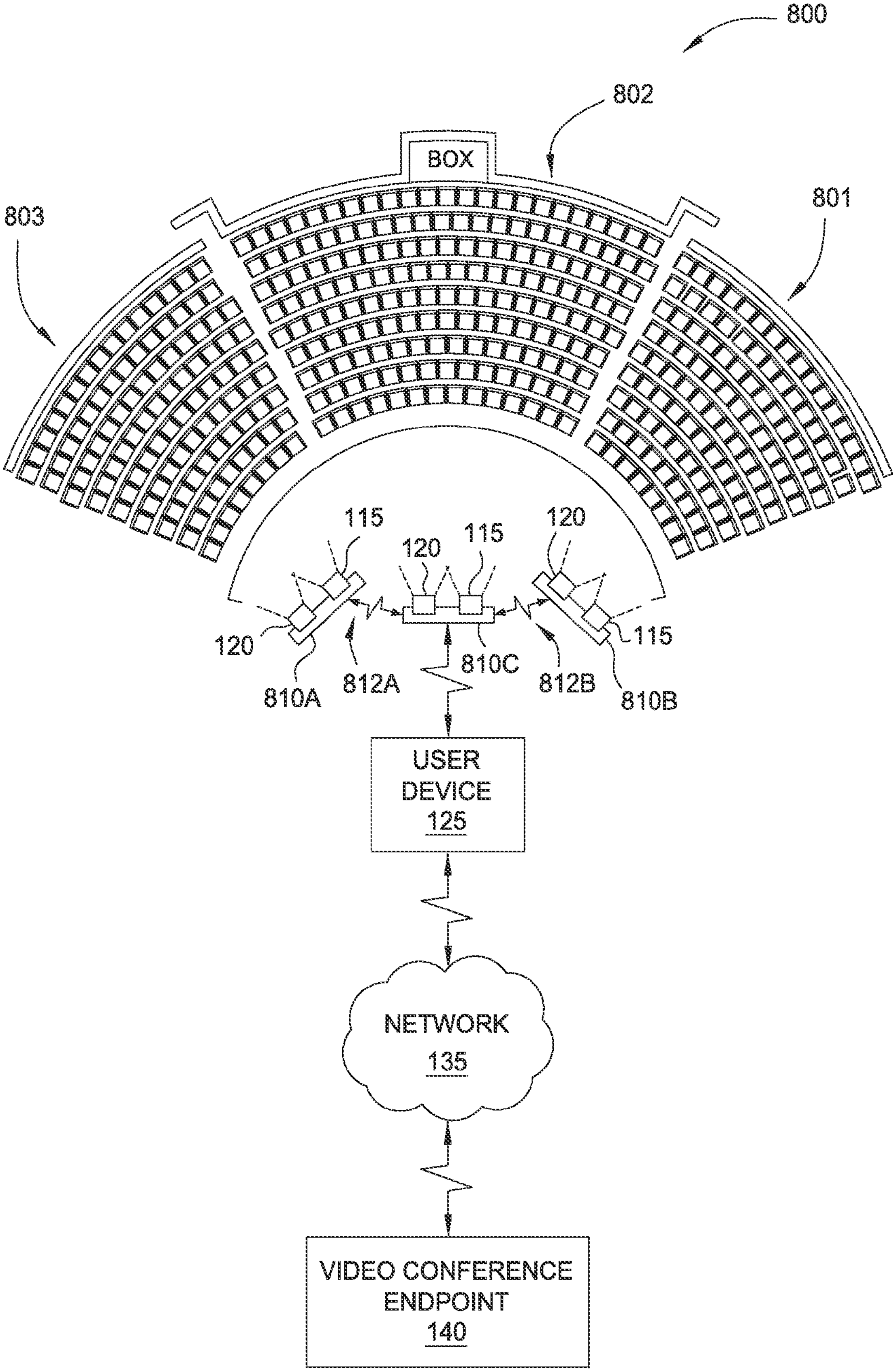

800
802
BOX
803
801
115
120
115
120
120
115
120
115
810A
812A
810C
812B
810B
USER DEVICE 125
NETWORK 135
VIDEO CONFERENCE ENDPOINT 140

FIG. 8

# SMART VIDEO CONFERENCING SYSTEM

## BACKGROUND

1. Field

The present invention relates to video cameras, and in particular to a smart video conferencing system that controls a video camera based on measures of activity within video data captured using another video camera.

2. Background

Video conferencing has become more popular in recent years, thanks in large part to proliferation of high speed Internet and price reductions in camera equipment. For example, dedicated video conferencing locations exist where rooms and technological resources are dedicated solely to the task of video conferencing. In particularly sophisticated environments that include multiple camera devices, server-side logic may be provided that is capable of dynamically switching between the video feeds of the various cameras when determining which video data to display at a remote video conferencing site. Additionally, many modern instant messaging software applications support voice and video chatting, where the participants can view each other while talking.

Generally, in capturing video data, video cameras are devices that can be configured to capture frames in a sequential manner using an image sensor. Additionally, a number of optimization operations can be performed on the captured frames in order to improve the quality of the video data. For instance, pixel correction operations can be performed on each captured frame, where bad pixel information is used to correct for hot or dead pixels. Additionally, auto focus operations can be performed, where a frame(s) is analyzed to determine whether the lens needs to be adjusted to achieve a more optimal focus. Upon determining that a lens adjustment is necessary, a feedback signal could be sent to motors or actuators to adjust the focal position of the lens. Additionally, color processing operations can be performed, where the frames are analyzed to determine if any color corrections are necessary. Such color corrections could include, for example, gamma correction, white balance correction and exposure correction.

Once any optimizations have been performed for the video data, the optimized video data can be encoded into a suitable format. Generally, the encoding format used can depend on the available network bandwidth and the application in question. For example, a dedicated video conference environment with a high bandwidth network connection could encode captured video data at a relatively high bit rate, while a video conference application on a mobile phone or tablet with a more limited network connection could encode the captured video data at a lower bit rate. The encoded video data can then be transmitted to the remote site for display via a communications network (e.g., the Internet).

While video conferencing technology is rapidly improving, it remains challenging to provided sophisticated video conferencing systems at relatively inexpensive prices. That is, while certain dedicated video conferencing environments provide many sophisticated features such as dynamically switching the displayed video stream between various captured video streams, such functionality currently comes at a substantial cost, in large part because such sophisticated setups require substantial computer hardware (e.g., multiple camera devices for capturing multiple different video feeds of the dedicated video conferencing environment, substantial network resources for transmitting multiple high-resolution video streams simultaneously, and server-side logic to select between the multiple different video feeds to determine what stream to display at the remote video conferencing site). As such, these sophisticated systems remain very expensive and priced beyond the practical reach of the average user.

Therefore, there is a need for a video conferencing system and method of using the same that solves the problems described above.

## SUMMARY

One embodiment presented in this disclosure provides a method of facilitating transmission of a video stream from a first video conferencing device to a remote video conferencing device. The method includes receiving, by the first video conferencing endpoint device, first video data captured from a first field of view of a physical environment. The video data includes a plurality of frames. The method also includes determining activity data for portions of the first video data across the plurality of frames. Additionally, the method includes generating, by a first video conferencing endpoint device, second video data from a second field of view of the physical environment, based on the determined activity data. The method further includes facilitating transmission of the video stream to the remote video conferencing device for display, where the video stream includes the generated second video data and audio data captured within the physical environment.

Embodiments of the disclosure may further provide a method of generating a video stream for use in a video conference, comprising receiving first video data captured from a first field of view of a physical environment, the first video data comprising a plurality of frames and determining activity data from portions of the first video data using information provided in the plurality of frames. Then generating second video data from a second field of view of the physical environment, based on the determined activity data, and generating a video stream that comprises the generated second video data and audio data captured within the physical environment.

Another embodiment presented in this disclosure provides a system for facilitating transmission of a video stream from a first video conferencing device to a remote video conferencing device. The system includes a first camera sensor configured to capture first video data comprising a plurality of frames from a first field of view of a physical environment. Additionally, the system includes a second camera sensor and a mounting structure capable of adjusting an orientation of the second camera sensor along one or more degrees of freedom. The system further includes control logic configured to determine activity data for portions of the video data across the plurality of frames and control movement of the mounting structure to adjust the orientation of the second camera along the one or more degrees of freedom, based on the determined activity data. The system also includes video processing logic configured to capture video data from a second field of view of the physical environment using the second camera sensor, encode the captured video data, facilitate transmission of the video stream to the remote video conferencing device for display, the video stream comprising the generated second video data and audio data captured within the physical environment.

Embodiments of the disclosure may further provide a system for generating a video stream for use in a video conference. The system includes a first camera sensor configured to capture first video data comprising a plurality of frames from a first field of view of a physical environment. Additionally, the system includes a second camera sensor and a mounting structure capable of adjusting an orientation of the second camera sensor along one or more degrees of freedom. The system further includes control logic configured to determine activity data for portions of the video data across the plurality of frames and control movement of the mounting structure to adjust the orientation of the second camera along the one or more degrees of freedom, based on the determined activity data. The system also includes video processing logic configured to capture video data from a second field of view of the physical environment using the second camera sensor, encode the captured video data, and generate a video stream comprising the generated second video data and audio data captured within the physical environment.

Embodiments of the disclosure may further provide a system for facilitating transmission of a video stream from a first video conferencing device to a remote video conferencing device. The system includes a camera sensor configured to capture first video data comprising a plurality of frames from a first field of view of a physical environment at a first resolution. The system also includes control logic configured to determine activity data for portions of the first video data across the plurality of frames, determine a portion of the captured first video data to extract, based on the determined activity data and extract the portion of the captured video data to create second video data. In such an embodiment, the second video data has a second resolution that is less than the first resolution of the captured video data. The system further includes video processing logic configured to facilitate transmission of the video stream to the remote video conferencing device for display. The video stream includes the generated second video data and audio data captured within the physical environment.

Embodiments of the disclosure may further provide a system for generating a video stream for use in a video conference. The system includes a camera sensor configured to capture first video data comprising a plurality of frames from a first field of view of a physical environment at a first resolution. The system also includes control logic configured to determine activity data for portions of the first video data across the plurality of frames, determine a portion of the captured first video data to extract, based on the determined activity data and extract the portion of the captured video data to create second video data. In such an embodiment, the second video data has a second resolution that is less than the first resolution of the captured video data. The system further includes video processing logic configured to generate a video stream that includes the generated second video data and audio data captured within the physical environment.

## BRIEF DESCRIPTION OF THE DRAWINGS

So that the manner in which the above-recited features of the present disclosure can be understood in detail, a more particular description of the disclosure, briefly summarized above, may be had by reference to embodiments, some of which are illustrated in the appended drawings. It is to be noted, however, that the appended drawings illustrate only typical embodiments of this disclosure and are therefore not to be considered limiting of its scope, for the disclosure may admit to other equally effective embodiments.

FIG. **1** is a diagram illustrating a video conferencing system for two video conferencing endpoints, according to one embodiment described herein.

FIG. **2** is a block diagram illustrating a video conferencing system for use as an endpoint to a video conference, according to one embodiment described herein.

FIG. **3** illustrates extracting video stream data from within a higher resolution video stream, according to one embodiment described herein.

FIG. **4** is a flow diagram illustrating a method for controlling a camera to capture video data based on activity detected within video data captured by another camera, according to one embodiment described herein.

FIG. **5** is a flow diagram illustrating a method for controlling a camera based on supplemental video streams and detected audio data, according to one embodiment described herein.

FIG. **6** is a flow diagram illustrating a method for creating a video stream by extracting a portion of a higher resolution video stream, according to one embodiment described herein.

FIG. **7** is a diagram illustrating a video conferencing environment that includes a video conferencing apparatus, according to one embodiment described herein.

FIG. **8** is a diagram illustrating a video conferencing environment that includes multiple video conferencing devices operating in a master-slave relationship, according to one embodiment described herein.

To facilitate understanding, identical reference numerals have been used, where possible, to designate identical elements that are common to the figures. It is contemplated that elements disclosed in one embodiment may be beneficially utilized on other embodiments without specific recitation.

## DETAILED DESCRIPTION

While much advancement has been made in video conferencing technology, it is still challenging to provide sophisticated video conferencing systems at relatively inexpensive prices. For instance, while dedicated video conferencing environments exist today that provide many sophisticated features such as multiple camera devices to capture multiple different video feeds of the dedicated video conferencing environment and server-side logic to select between the multiple different video feeds to determine what stream to display at the remote video conferencing site, such dedicated video conferencing systems remain very expensive and may be priced beyond the practical reach of the average user. Moreover, such sophisticated setups typically require much more network bandwidth, as multiple video streams are transmitted to an intermediary server that then selects between the multiple video streams.

On the other hand, inexpensive video conferencing systems exist but are predominantly limited to transmitting only a single video stream across the network. While this lowers the network requirements for using such video conferencing systems, it also limits what can be displayed at the remote video conferencing site to the single video stream. Thus, for video captured in a conference room environment, the remote site may simply see an overall view of the conference room environment, and any adjustments to the camera (e.g., pan, tilt and zoom operations) must be made manually by a user.

As such, embodiments provide a video conferencing endpoint system capable of manipulating a field of view of captured video content that is transmitted to a remote video conferencing endpoint device, based on measures of activity between frames of captured video content. For instance, embodiments may receive captured video data having a first field of view of a physical environment. Generally, the captured video data includes a sequential plurality of frames captured using a camera sensor. In one embodiment, the video data is captured using a wide angle camera sensor.

Embodiments may then analyze the captured video data to determine a plurality of measures of activity for portions of the video data across the plurality of frames. Generally, the measures of activity correspond to types of activities that are desirable to include in the video stream transmitted to the remote video conferencing site. For example, embodiments could analyze the captured video data to determine which of a plurality of users depicted within the video data is currently speaking. In doing so, embodiments could first perform a facial recognition analysis on the frames of the video data and could then determine a measure of motion for each detected face within the frames of the video data. Embodiments could then determine which measure of motion is most indicative of a user speaking.

Embodiments could then use the determined measures of activity to generate stream video data having a second field of view of the physical environment. For instance, embodiments could control the movement of a second camera device to capture stream video data that is substantially centered on a physical entity within the physical environment, based on the determined measures of activity. As an example, logic for the video conferencing system could determine which user depicted within the captured video data is currently speaking and could control the movement of the second camera device (e.g., pan, tilt and zoom of the second camera device) to capture stream video data that is substantially centered on the user who is currently speaking. The stream video data could then be transmitted over a network to a remote video conferencing device for display. Doing so provides an intelligent video conferencing system that is capable of dynamically adjusting the video stream that is transmitted to the remote video conferencing device.

An example of such a system is shown in FIG. **1**, which is a diagram illustrating a video conferencing system for two video conferencing endpoints, according to one embodiment described herein. As shown, the system **100** includes a video conferencing endpoint **110** and a video conferencing endpoint **140**, interconnected via a network **135**. The network **135** generally represents any data communications network suitable for the transmission of video and audio data (e.g., the Internet). In some configurations, each of the video conferencing endpoints **140** includes one or more display devices for at least displaying received video and audio data and video and audio capture devices (e.g., pan and tilt camera **120**) for capturing video data to send to the other video conferencing endpoints **110**, **140**. For example, the video conferencing endpoint **140** could represent a video conferencing software application (e.g., Microsoft® Skype®) executing on a mobile computing device that transmits captured video and audio data across the network **135** for display at the video conferencing endpoint **110** and that displays video and audio data received from the video conferencing endpoint **110**. As another example, the video conferencing endpoint **140** could represent a dedicated video conferencing environment in which multiple camera devices are permanently installed. More generally, the video conferencing endpoint **140** represents any device(s) that are suitable to participate in a video conference, or even just displaying the received video and audio data.

Of note, while numerous examples are provided herein involving capturing video stream data (e.g., using one or more camera devices) and transmitting the video stream data as part of a video conference, one of ordinary skill in the art will understand that such video stream data may include one or more captured audio streams as well (e.g., captured using one or more microphone devices within the physical environment). As such, while a particular example may be described as transmitting video data, a video stream and/or video stream data to a remote video conferencing site, it is explicitly contemplated that captured audio data can be transmitted as well using any suitable technique for representing and transmitting audio data.

As shown, the video conferencing endpoint **110** includes a wide angle camera device **115**, a pan and tilt camera device **120**, a control device **130**, one or more microphones **131** and a user system **125**. Generally, the wide angle camera device **115** is configured to capture a video stream of the physical environment, and preferably is positioned so that all users within the physical environment are depicted within the captured video stream. The pan and tilt camera device **120** is capable of adjusting the orientation of a camera sensor within the device **120**, across multiple degrees of freedom. Additionally, the pan and tilt camera device **120** may be capable of zoom functionality, e.g., hardware zoom or software zoom. While the system **100**, shown in FIG. **1**, illustrates a configuration in which the wide angle camera device **115**, the pan and tilt camera device **120** and the control device **130** are distinct and separate components, this configuration is not intended to limit the scope of the disclosure provided herein, since other configurations or groupings of these components are also envisioned. In one example, the system **100** may include a single enclosure, such as a video conferencing apparatus **720** (FIG. **7**) discussed below, that includes the control device **130** and one or more camera devices (e.g., wide angle camera device **115** and one or more pan and tilt camera device **120**). In another example, the system **100** may include a single enclosure that includes the control device **130**, one or more microphones **131** and the wide angle camera device **115**. In yet another example, the system **100** may include a single enclosure that includes the control device **130**, one or more microphones **131**, the wide angle camera device **115** and the tilt camera device **120**. In some cases, it is desirable to position the camera devices so that they face the same direction (e.g., front facing) and are oriented so that their fields-of-view completely overlap.

Generally, the user device **125** represents any computing device capable of transmitting a video stream to a remote video conferencing device (e.g., video conferencing endpoint **140**) over the network **135**. Examples of the user device **125** include personal computing devices, tablet computing devices, mobile devices and so on. Moreover, the user device **125** can execute one or more video conferencing software applications capable of receiving video data from the control device **130** (e.g., captured using the pan and tile camera device **120**) and transmitting the captured video data to the video conferencing endpoint **140**, via the network **135**. Examples of such video conferencing applications include, without limitation, Microsoft® Skype® and Apple® FaceTime®. More generally, however, any video conferencing application capable of receiving video data and transmitting the video data to a remote site can be used, consistent with the functionality described herein. Additionally, the user device **125** may display video data captured at and received from the remote video conferencing endpoint **140**, e.g., on a native display device of the user device **125** or on a separate display device (not shown) within the physical environment.

The control device **130** is generally configured to adjust and orientation of the pan and tilt camera device **120**, based on detected measures of activity within a video stream captured by the wide angle camera device **115**. For example, the control device **130** could analyze the video stream captured by the wide angle camera device **115** to determine which user depicted within the video data is currently speaking. The control device **130** could then adjust the orientation of the pan and tilt camera device **120**, so that the video stream captured by the pan and tilt camera device **120** is substantially centered on the determined user that is currently speaking within the physical environment. In one embodiment, the control device **130** can also adjust the zoom level of the pan and tilt camera device **120**, so that the user who is currently speaking is depicted at a predetermined size within the video stream captured by the pan and tilt camera device **120**. For example, the control device **130** could adjust the zoom level such that the user occupies 70% of the captured video frames.

The control device **130** could then transmit the video data captured using the pan and tilt camera **120** and the microphone **131** to the user device **125**. Upon receiving the video data, the user device **125** could transmit the video data to the video conferencing endpoint **140** for display, using the network **135**. Advantageously, the control device **130** provides intelligent camera control for video conferencing applications on the user device **125**, without requiring the transmission of multiple video streams across the network **135**.

FIG. **2** is a block diagram illustrating a video conferencing system for use as an endpoint to a video conference, according to one embodiment described herein. For example, the system **200** could be used as either the video conferencing endpoint **110** and/or **140** and may contain one or more of the related components shown in FIG. **1**. As shown, the video conferencing system **200** includes a wide angle camera sensor **210**, pan and tilt camera sensor **220**, camera controller component **230**, video streaming component **240**, and microphone sensors **260**. Generally, the wide angle camera sensor **210** represents a camera having a wide angle lens that is configured to capture video data of the physical environment. Preferably, such a wide angle camera sensor **210** is positioned so that all or substantially all users within the physical environment, and in front of the camera, are depicted within the captured video. The wide angle camera sensor **210** may have a field of view that has a viewing angle that is between about 70 degrees and about 180 degrees, such as a viewing angle of about 130 degrees. Additionally, the wide angle camera sensor **210** is capable of providing the captured video data to the camera controller component **230**. As an example, the camera controller component **230** could use such data to control movement of the pan and tilt camera sensor **220**, create the video stream to transmit to a remote video conferencing endpoint.

The pan and tilt camera sensor **220** represents a camera sensor affixed to a mounting structure that is capable of moving in one or more degrees of freedom. For example, the mounting structure may be capable of moving such that an orientation of the pan and tilt camera sensor **220** changes in the horizontal and/or vertical directions. Additionally, the pan and tilt camera sensor **220** can be capable of zoom functionality (e.g., hardware zoom, software zoom, etc.). The pan and tilt camera sensor **220** is capable of providing the captured video data to the camera controller component **230**. As an example, the camera controller component **230** could use the data from the pan and tilt camera sensor **220** in creating the video stream to transmit to the remote video conferencing endpoint. The pan and tilt camera sensor **220** may have a field of view that has a viewing angle that is between about 5 degrees and about 80 degrees, such as a viewing angle of about 50 degrees.

The camera controller component **230** is generally configured to adjust the orientation of the pan and tilt camera sensor **220**, based on detected measures of activity within video data captured by the wide angle camera sensor **210**. For example, the camera controller component **230** could analyze the received video data captured using the wide angle camera sensor **210** to detect which user within the physical environment is currently speaking. The camera controller component **230** could then adjust the orientation of the pan and tilt camera sensor **220**, such that the user currently speaking is substantially centered within video data captured using the pan and tilt camera sensor **220**. As another example, the camera controller component **230** could analyze the video data captured using the wide angle camera sensor **210** and could determine that a particular user has moved to a predefined location within the physical environment (e.g., within a predefined distance of a whiteboard within the physical environment). The camera controller component **230** could then adjust the orientation of the pan and tilt camera sensor **220**, such that the particular user is substantially centered or desirably positioned within the video data captured using the pan and tilt camera sensor **220**.

Typically, the camera controller component **230** is a general use computing device that includes a processor **218**, a memory unit **214**, input/output (I/O) devices **216** and a communications device **250**. The memory unit **214** is generally included to be representative of a random access memory. The memory unit **214** represents any memory sufficiently large to hold the necessary programs and data structures. Memory unit **214** could be one or a combination of memory devices, including Random Access Memory, nonvolatile or backup memory (e.g., programmable or Flash memories, read-only memories, etc.). In addition, memory unit **214** may be considered to include memory physically located elsewhere, for example, on another computer communicatively coupled to the camera controller device. Illustratively, the memory unit **214** includes an operating system **219**, one or more software applications **215**, stored media data **217** and a model structure **213**. Generally, the model structure **213** describes attributes of the physical environment in which the video conferencing system **200** is located (e.g., a make-up and arrangement of physical entities within the physical environment). Examples of operating system **219** include UNIX, a version of the Microsoft Windows® operating system, and distributions of the Linux® operating system. Additional examples of operating system **219** include custom operating systems for gaming consoles, including the custom operating systems for systems such as the Nintendo DS® and Sony PSP®.

Processor **218** may be a hardware unit or combination of hardware units capable of executing software applications and processing data. In some configurations, the processor **218** includes a central processing unit (CPU), a digital signal processor (DSP), an application-specific integrated circuit (ASIC), and/or a combination of such units. Generally, the processor **218** retrieves and executes programming instructions stored in the memory unit **214**. Processor **218** is included to be representative of a single CPU, multiple CPUs, a single CPU having multiple processing cores, GPUs having multiple execution paths, and the like. For example, the processor **218** could execute the one or more software applications **215** and process the stored media data **217**, which are each included within memory unit **214**.

The video streaming component **240** is generally configured to facilitate the transfer of video data (e.g., video data captured using the pan and tilt camera sensor **220**) to a remote video conferencing device for display (e.g., using a communications device **250** such as a network interface controller). For example, the video streaming component **240** can encode the captured video data in an encoding format and at a bit rate suitable for the video conferencing session and deliver the video data to the user device **125** via a wireless (e.g., WiFi, Bluetooth®, etc.) or wired connection (e.g., USB connection). The user device **125** could then transmit the video data to a remote video conferencing device using conventional communication devices and protocols (e.g., network interface card, Ethernet card, modem, wireless network hardware or other conventional computing device communication hardware). In some configurations, the video streaming component **240** provides the video data to a video streaming application on the user device **125**, such as, without limitation, Microsoft® Skype® and Apple® FaceTime®, using communication device **250**

The encoding format used by the video streaming component **240** could be determined based on the capabilities or specified preferences of the user device **125**. One example of a suitable video encoding format is H.264 (and variants such as H.264/AVC, H.264 High Profile and H.264 SVC). More generally, however, any suitable encoding format can be used, consistent with the present disclosure. For example, a particular video streaming application on the user device **125** could be configured to transmit streaming video data in a particular encoding format, and thus the video streaming component **240** could be configured to encode the generated video data using the particular encoding format for delivery to the streaming application on the user device. Additionally, the remote video conferencing device may be capable of receiving the encoded format from the user device and output the received data on a display. In one embodiment, the encoding format can be specified by a user of the video conferencing system **200**. Moreover, in some instances, the user device **125** (or a video streaming application executing on the user device) can transcode the received video data to another format, before transmitting the video data to the remote video conferencing endpoint device.

In one embodiment, the camera controller component **230** can adjust the orientation of the pan and tilt camera sensor **220** in order to intelligently capture video data within the physical environment and the camera controller component **230** can then transmit the captured video data to the video conferencing application on the user device (e.g., over a wireless network connection, over a wired connection, etc.). The video conferencing application could then perform any needed encoding operations on the received video data (e.g., transcoding the received video data from a first format to another format used by the video conferencing application) and could then transmit the video data to a remote video conferencing device for display. Advantageously, in such an embodiment the camera controller component **230** provides intelligent video conferencing services for existing video conferencing applications of the user device, thereby improving the performance of the video conferencing applications on the user device without any modifications to the video conferencing applications.

In one embodiment, the camera controller component **230** is configured to use the microphone sensors **260** in order to determine an orientation for the pan and tilt camera sensor **220**. For instance, the microphone sensors **260** could represent a microphone array (e.g., two or more microphone sensor devices operating in tandem) and the camera controller component **230** could include logic to determine a direction from which particular audio content was received by the microphone sensors **260** within the microphone array. As an example, the camera controller component **230** could analyze audio data collected by the microphone sensors **260** within the microphone array and could determine that a portion of the audio data matches a predefined signature corresponding to user speech data. The camera controller component **230** could then determine a direction from which the user speech data originated, relative to a physical position of the microphone sensors **260**. The camera controller component **230** could then use the determined direction to orient the pan and tilt camera sensor **220**, such that the pan and tilt camera sensor **220** is oriented in substantially the determined direction from which the user speech data originated. By doing so, the camera controller component **230** could capture video data of the user who is currently speaking within the physical environment. As discussed above, in doing so, the camera controller component **230** can also alter a zoom level of the pan and tilt camera sensor **220**, so that the user that is speaking is predominantly featured within the captured video data. For example, the camera controller component **230** could alter the zoom of the pan and tilt camera sensor **220** until the speaking user occupies a predefined portion of the captured frames of video data.

In one embodiment, the camera controller component **230** is configured to use a single high-resolution camera to capture a video stream that is at least partially transmitted to the remote video conferencing device. An example of this is shown in FIG. **3**, which illustrates extracting a video data from within a higher resolution video stream, according to one embodiment described herein. As shown, the screenshot illustrates a plurality of users within a video conferencing environment. In the depicted example, the user **305** is currently speaking. As discussed above, the camera controller component **230** could perform an analysis of the captured video data to determine a physical entity (e.g., a particular user) to be the focus of the stream video data transmitted to the remote video conferencing device. For instance, the camera controller component **230** could determine that the frames should focus on the user **305** who is currently speaking within the physical environment.

Upon determining that the user **305** is currently speaking, the camera controller component **230** could determine a portion of the captured video data to extract as stream video data to be transmitted to the remote video conferencing device. For example, in the illustration **300**, the camera controller component **230** has determined the region **310** around the user **305**, and the controller component **230** could extract the region **310** from each frame of the captured video data to create a stream of video data. In doing so, the camera controller component **230** can dynamically adjust the region **310** for each frame of the captured video, so that the user **305** is substantially centered within each frame of the stream video data. Doing so allows a single camera to be used for both the detection of activity within the physical environment and the capture of video data to be streamed to the remote video conferencing device.

Generally, as each frame of the stream video data is extracted from a portion of a corresponding frame of the captured video data, the stream video data is a lower resolution video stream than the captured video data. For example, and without limitation, the captured video data could have a resolution of 3840×2160 pixels, while the stream video data extracted from the captured video data could have a resolution of 1024×768 pixels. Such an embodiment may be preferable, for instance, because many modern devices have the computing resources and network resources to support the transmission and display of video data with a resolution of 1024×768 pixels, while substantially fewer devices today may be capable of supporting (i.e., with sufficient computing and network resources) a video streaming having a resolution of 3840×2160 pixels. In one example, the stream video data is sent to the external endpoint from the user device **125** in a 1080p resolution (or other desirable resolution) using typical computing and network resources. Likewise, the stream video data can be sent to the user device **125** in a 1080p resolution (or other desirable resolution), via the video conferencing system (e.g., system **200**), so that the user device

**125** can then send the stream video data to the external endpoint. More generally, however, any suitable resolutions for the captured video data and the stream video data can be used, consistent with the present disclosure.

FIG. **4** is a flow diagram illustrating a method for controlling a camera to capture video data based on activity detected within video data captured by another camera, according to one embodiment described herein. As shown, the method **400** begins at block **410**, where a wide-angle camera sensor captures a wide-angle video stream of a video conferencing environment. The camera controller component **230** analyzes the wide-angle video stream to detect a region of activity within the frames of the video stream (block **415**). As discussed above, examples of such activity include (without limitation) a user currently speaking and a user moving to a predefined location within the physical environment (e.g., a whiteboard).

The camera controller component **230** then manipulates a second camera device (e.g., pan and tilt camera sensor **220**) to capture a video stream corresponding to the detected region of activity (block **420**). As an example, upon determining that a particular user is currently speaking based on an analysis of the captured wide-angle video stream, the camera controller component **230** could manipulate the position of the second camera device such that the second camera device is substantially oriented in the direction of the particular user to collect video data of the user. A video streaming component **240** can then receive the video data captured using the second camera device and can facilitate the transfer of the captured video data to a remote video conferencing endpoint for display (block **425**), and the method **400** ends. For example, the video streaming component **240** could provide the captured video data to a streaming application on a user device **125**, which in turn transmits the video data to the remote videoconferencing endpoint (e.g., via the Internet). Advantageously, doing so provides an intelligent video conferencing system that is capable of providing a video stream that selectively focuses on physical elements within the video conferencing environment (e.g., a user who is currently speaking), without requiring multiple, distinct video streams to be transmitted across the network from the video conferencing system (e.g., to an intermediary server system for processing), thereby reducing the computing and network resources used in providing video conferencing services.

FIG. **5** is a flow diagram illustrating a method for controlling a camera based on supplemental video streams and detected audio data, according to one embodiment described herein. As shown, the method **500** begins at block **510**, where the camera controller component **230** analyzes one or more supplemental video streams to determine a respective one or more measures of activity for regions of the supplemental video streams. For example, one of the supplemental video streams could be the video stream captured using a wide-angle camera sensor. In some environments, additional camera sensors may be provided, such as a camera sensor oriented substantially in the direction of a predefined area of interest (e.g., a whiteboard within the video conferencing environment). The camera controller component **230** can analyze the frames of each of these video streams in order to determine an orientation for a pan and tilt camera device that captures video data to be transmitted to a remote video conferencing device.

Additionally, the camera controller component **230** collects audio data using microphones sensors within the physical environment (block **515**) and determines a direction from which at least a portion of the audio data originated (block **520**). For instance, as discussed above, and microphone array could be used to capture the audio data and, upon identifying that a portion of the audio data matches a predefined profile for user speech, the camera controller component **230** could use the data collected from the microphone sensors in the microphone array to determine a direction from which the user speech originated.

The camera controller component **230** then manipulates an orientation of a camera device (e.g., a pan and tilt camera sensor **220**) based on the determined measures of activity, the determined direction from which at least a portion of the audio data originated, and a mapping structure describing a layout of the physical environment (block **525**). For example, the mapping structure could specify a location of physical elements within the physical environment, such as chairs in which the users may be seated, and the camera controller component **230** could use this mapping structure to more accurately orient the camera device in a direction of interest. Additionally, such a mapping structure could specify predefined areas of interest within the physical environment, such as the location of a whiteboard relative to a physical position of the camera devices. If the camera controller component **230** then detects activity within frames of the supplemental video stream(s) that is indicative of a user moving to the predefined area of interest within the physical environment, the camera controller component **230** could manipulate the orientation of the camera device to face (and capture video data of) the predefined area of interest.

The video streaming component **240** receives the video data captured using the manipulated camera device and encodes the captured video data in a suitable format (block **530**). The video streaming component **240** then facilitates the transmission of the encoded to video data to a remote video conferencing endpoint for display (block **535**), and the method **500** ends. For example, the video streaming component **240** transmits the encoded video data (e.g., using a communications device **250**, such as a BlueTooth® transceiver or a wired connection) to a user device **125**, and a video streaming application on the user device **125** in turn transmits the encoded video data (e.g., using a network interface controller or wireless interface controller) to the remote video conferencing endpoint device for display.

FIG. **6** is a flow diagram illustrating a method for creating a video stream by extracting a portion of a higher resolution video stream, according to one embodiment described herein. As shown, the method **600** begins at block **610** where a high-resolution camera captures video data of a physical environment. The camera controller component **230** then performs a facial recognition analysis to identify a plurality of video conferencing participants within the frames of the captured video data (block **615**). The camera controller component **230** also determines a measure of motion for each identified participant across the frames of the video data (block **620**). In doing so, the camera controller component **230** can determine a region of the frames that corresponds to each participant's mouth and can restrict the motion determination to this region. That is, the camera controller component **230** can determine which of the participants is currently speaking within the physical environment, and may ignore other forms of motion such as one of the participants nodding or scratching his head.

The camera controller component **230** then selects one of the identified participants, based on the determined measures of motion (block **625**). For example, the camera controller component **230** could select the participant whose movement most closely matches a predefined movement profile indicative of user speech. Of note, may not be the user with the most motion from frame to frame, as the camera controller component **230** may select the user who is determined to be speaking as opposed to a user who is simply moving throughout the physical environment (e.g., the user arriving late to the video conference and in the process of sitting down).

Upon selecting one of the participants, the camera controller component **230** creates a second video stream having a lower resolution than the captured video stream, by extracting a portion of video data from each frame of the captured video stream that includes the selected participant (block **630**). In doing so, the camera controller component **230** essentially creates a virtual camera focused on the selected participant. Generally, the size of the extracted portion of video data corresponds to a desired resolution of the video stream for the video conference. For example, if the video conference is configured to use video data having a resolution of 1024×768 pixels, the camera controller component **230** could extract a 1024×768-sized portion from each frame of the captured video. A video streaming component **240** facilitates the transfer of the created second video stream to a remote video conferencing site for display (block **635**), and the method **600** ends. In one configuration, as discussed above, the video streaming component **240** transmits the video data to a user device **125**, which in turn transmits the video data to the remote video conferencing site for display. Advantageously, doing so allows a single camera to be used for both the motion detection analysis and for the capture of the video data to be streamed to the remote video conferencing site.

In one embodiment, multiple camera devices are provided to capture multiple video streams of the physical environment, and the generated video stream that is transmitted to the remote video conferencing device is made up of multiple captured video streams. Moreover, these camera devices as well as control logic for controlling the movement of the camera devices can be provided within a single enclosure. An example of this is shown in FIG. **7**, which is a block diagram illustrating a video conferencing environment that includes a video conferencing apparatus, according to one embodiment described herein. As shown, the diagram includes a video conferencing environment **700**, which includes a video conferencing apparatus **720**, a whiteboard **714**, user participants **718** and a currently speaking user **716**. The video conferencing apparatus **720** includes a wide angle camera device **115** having a field of view **722**, pan and tilt camera devices **120**A and **120**B having fields of view **724** and **726**, respectively, and a control device **130**. In one configuration, as shown in FIG. **7**, the pan and tilt camera devices **120**A is positioned and/or oriented to view the whiteboard, the pan and tilt camera devices **120**B is positioned and/or oriented to view a different portion of the video conferencing environment **700** and the wide angle camera device **115** is positioned and/or oriented to view at least a portion of the areas viewed by both of the pan and tilt camera devices **120**A and **120**B.

As discussed above, the control device **130** can analyze the video data captured by the wide angle camera device **115** to determine how to manipulate the pan and tilt camera devices **120**A and **120**B. For purposes of the depicted example, assume that the control device **130** has analyzed the video stream from the wide angle camera device **115** and has determined that the user **716** is currently speaking, and based on this, has manipulated the orientation of the pan and tilt camera device **120**B so that the user **716** is substantially centered within the video frames captured by the pan and tilt camera device **120**B.

The control device **130** can generate a video stream for transmission to the remote video conference endpoint **140**. As discussed above, the control device **130** could transmit video data captured using the pan and tilt camera device **120**B to the user device **125**, which then transmits the video data to the remote video conference endpoint **140** for display, using the network **135**. In the depicted embodiment, the control device **130** is configured to generate a composite video stream that includes video data captured from all of the pan and tilt camera devices **120**A, **1208** and wide angle camera device **115**. The control device **130** can then provide the generated composite video stream to the user device **125**, for transmission to the video conference endpoint **140** using network **135**.

Interface **701** illustrates an example of a rendering of the output of a composite video stream, where portion **702** corresponds to video data captured using the wide angle camera device **115**, portion **704** corresponds to video data captured using pan and tilt camera device **120**A, and portion **706** corresponds to video data captured using pan and tilt camera device **120**B. However, in some configurations of the interface **701**, the portion **704** may be formed from a portion of the video data captured using the wide angle camera device **115** (e.g., a sub-region of portion **702**), as similarly discussed above in conjunction with FIG. **3**. Rendering of the composite video stream may be performed by simultaneously displaying each of the video data elements found in the composite video stream on a display device that is part of the video conference endpoint **140**. Of course, the interface **701** is provided without limitation and for illustrative purposes only, and of course any number of different arrangements and combinations of video streams can be used, consistent with the present disclosure. Advantageously, by generating the video stream at the control device **130**, embodiments limit the data transmitted over the network **135** to only a single video stream from the environment **700** to the video conference endpoint **140**. That is, while the interface **701** includes multiple captured video streams, only a single composite video stream is provided by the control device **130** to the user device **125** and is subsequently transmitted across the network **135** to the video conference endpoint **140**, thereby providing a more intelligent and sophisticated video conferencing system while reducing the needed computing and networking resources, relative to conventional techniques.

In addition to a single enclosure that includes multiple camera devices (e.g., the enclosure **720** that includes camera devices **115**, **120**A and **120**B and in some configurations a microphone (not shown)), in one embodiment multiple video conferencing enclosures are configured to operate in tandem in a video conferencing environment to provide a single video stream to a remote endpoint device. Such an embodiment may be preferable, for instance, when capturing video streams of a large physical environment such as an auditorium, where a single camera device may not have sufficient field of view to capture video data that includes all of the participants to the video conference. An example of this is shown in FIG. **8**, which is a diagram illustrating a video conferencing environment that includes multiple video conferencing devices operating in a master-slave relationship, according to one embodiment described herein. As shown, the illustration **800** includes auditorium seating sections **801**, **802** and **803**, and video conferencing enclosures **810**A-C. Each of the video conferencing enclosures **810**A-C includes a respective pan and tilt camera sensor **120** and a respective wide angle camera sensor **115**. Additionally, each of the video conferencing enclosures **810**A-C may include respective control logic **130**. In one example, each of the video conferencing enclosures **810**A-C include a video conferencing apparatus **720**, which is discussed above.

As shown, the video conferencing enclosures **810**A and **810**C are connected by communication link **812**A (e.g., a wired communication link, a wireless communication link, etc.) and the video conferencing enclosures **810**B and **810**C are connected by communication link **812**B. In the depicted embodiment, the video conferencing enclosure **810**C is configured to act in a master device role, while the video conferencing enclosures **810**A and **810**B are configured to act in a slave device role. That is, the video conferencing enclosures **810**A and **810**B are configured to provide video stream data captured using the respective camera sensors **115** and/or **120**, and the video conferencing enclosure **810**C is configured to generate a video stream to transmit to a remote video conferencing endpoint device, based on the received video streams. Additionally, the video conferencing enclosure **810**C can further generate the video stream based on video data captured using the camera sensors **115** and **120** on the video conferencing enclosure **810**C. For example, control logic for the video conferencing enclosure **810**C can analyze the received video data received from the other video conferencing enclosures, as well as the video data captured using the camera sensors **115** and **120** on the video conferencing enclosure **810**C, to determine which video stream(s) to include in the generated video stream. As discussed above, such a generated video stream can be composed of video data captured using a single camera device or can be a composite video stream that includes video data captured using multiple camera devices (e.g., multiple camera devices within a single video conferencing apparatus, multiple camera devices from multiple video conferencing apparatuses, etc.).

The generated video stream is then transmitted to the user device **125** (e.g., a personal computing device executing a video streaming software application, such as Microsoft® Skype®). Upon receiving the generated video stream, the user device **125** transmits the video data to the video conference endpoint device **140** (e.g., a remote user device that is also executing a respective video streaming software application) using the network **135**. Doing so allows the video conferencing techniques disclosed herein to be extended to any sized physical environment through the use of additional video conferencing enclosures.

In the preceding, reference is made to embodiments presented in this disclosure. However, the scope of the present disclosure is not limited to specific described embodiments. Instead, any combination of the described features and elements, whether related to different embodiments or not, is contemplated to implement and practice contemplated embodiments. Furthermore, although embodiments disclosed herein may achieve advantages over other possible solutions or over the prior art, whether or not a particular advantage is achieved by a given embodiment is not limiting of the scope of the present disclosure. Thus, the preceding aspects, features, embodiments and advantages are merely illustrative and are not considered elements or limitations of the appended claims except where explicitly recited in a claim(s).

As will be appreciated by one skilled in the art, the embodiments disclosed herein may be embodied as a system, method or computer program product. Accordingly, aspects may take the form of an entirely hardware embodiment, an entirely software embodiment (including firmware, resident software, micro-code, etc.) or an embodiment combining software and hardware aspects that may all generally be referred to herein as a "circuit," "module" or "system." Furthermore, aspects may take the form of a computer program product embodied in one or more computer readable medium(s) having computer readable program code embodied thereon.

Any combination of one or more computer readable medium(s) may be utilized. The computer readable medium may be a computer readable signal medium or a computer readable storage medium. A computer readable storage medium may be, for example, but not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples (a non-exhaustive list) of the computer readable storage medium would include the following: an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the foregoing. In the context of this document, a computer readable storage medium is any tangible medium that can contain, or store a program for use by or in connection with an instruction execution system, apparatus or device.

A computer readable signal medium may include a propagated data signal with computer readable program code embodied therein, for example, in baseband or as part of a carrier wave. Such a propagated signal may take any of a variety of forms, including, but not limited to, electro-magnetic, optical, or any suitable combination thereof. A computer readable signal medium may be any computer readable medium that is not a computer readable storage medium and that can communicate, propagate, or transport a program for use by or in connection with an instruction execution system, apparatus, or device.

Program code embodied on a computer readable medium may be transmitted using any appropriate medium, including but not limited to wireless, wireline, optical fiber cable, RF, etc., or any suitable combination of the foregoing.

Computer program code for carrying out operations for aspects of the present disclosure may be written in any combination of one or more programming languages, including an object oriented programming language such as Java, Smalltalk, C++ or the like and conventional procedural programming languages, such as the "C" programming language or similar programming languages. The program code may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider).

Aspects of the present disclosure are described herein with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems) and computer program products according to embodiments presented in this disclosure. It will be understood that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the flowchart illustrations and/or block diagrams, can be implemented by computer program instructions. These computer program instructions may be provided to a processor of a general purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks.

These computer program instructions may also be stored in a computer readable medium that can direct a computer, other programmable data processing apparatus, or other devices to function in a particular manner, such that the instructions stored in the computer readable medium produce an article of manufacture including instructions which implement the function/act specified in the flowchart and/or block diagram block or blocks.

The computer program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other devices to cause a series of operational steps to be performed on the computer, other programmable apparatus or other devices to produce a computer implemented process such that the instructions which execute on the computer or other programmable apparatus provide processes for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks.

The flowchart and block diagrams in the Figures illustrate the architecture, functionality and operation of possible implementations of systems, methods and computer program products according to various embodiments. In this regard, each block in the flowchart or block diagrams may represent a module, segment or portion of code, which comprises one or more executable instructions for implementing the specified logical function(s). It should also be noted that, in some alternative implementations, the functions noted in the block may occur out of the order noted in the figures. For example, two blocks shown in succession may, in fact, be executed substantially concurrently, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts, or combinations of special purpose hardware and computer instructions.

In view of the foregoing, the scope of the present disclosure is determined by the claims that follow.

I claim:

**1**. A method of generating a video stream for use in a video conference, comprising:

receiving, by a first video conferencing endpoint device, first video data captured from a first field of view of a physical environment, the first video data comprising a plurality of frames;

determining activity data from portions of the first video data using information provided in the plurality of frames;

generating, by the first video conferencing endpoint device, second video data from a second field of view of the physical environment, based on the determined activity data;

generating a video stream that comprises the generated second video data and audio data captured within the physical environment; and

transmitting the video stream to a video conferencing application executing on a user device, wherein the video conferencing application is configured to process the video stream as an input video stream to facilitate the transmission of the video stream to a remote video conferencing device for display.

**2**. The method of claim **1**, wherein generating the second video data from the second field of view of the physical environment further comprises controlling movement of a controlled camera device to capture the second video data, and wherein the received first video data is captured using a wide angle camera device, which is distinct from the controlled camera device.

**3**. The method of claim **2**, wherein determining the activity data for portions of the first video data using information provided in the plurality of frames further comprises:

performing a facial detection analysis to detect a plurality of user faces within the first video data; and

determining a measure of motion for each of the detected plurality of user faces using information provided in the plurality of frames of the first video data.

**4**. The method of claim **3**, wherein generating the second video data from the second field of view of the physical environment further comprises:

selecting one of the plurality of user faces having a corresponding determined measure of motion that is indicative of user speech; and

determining an orientation of the camera device for capturing video data substantially centered on the selected user face, and

wherein controlling the movement of the camera device to capture the second video data further comprises controlling the movement of the camera device to match the determined orientation.

**5**. The method of claim **4**, wherein determining an orientation of the camera device for capturing the video data substantially centered on the selected user face further comprises:

identifying a physical entity corresponding to the selected user face by accessing a model structure describing an attribute of the physical environment; and

determining a direction of the identified physical entity, relative to a physical position of the camera device within the physical environment.

**6**. The method of claim **4**, wherein generating the second video data from the second field of view of the physical environment further comprises:

collecting audio data from the physical environment using two or more microphone sensors;

identifying user speech within the collected audio data; and

determining a direction from which the identified user speech originates, relative to a physical position of the two or more microphone sensors, and

wherein determining the orientation of the camera device for capturing the video data substantially centered on the selected user face is further based on the determined direction from which the identified user speech originates.

**7**. The method of claim **1**, wherein the received first video data is captured at a first resolution, and wherein generating the second video data from the second field of view of the physical environment further comprises:

extracting a portion of the first video data to create the second video data, wherein the second video data has a second resolution that is less than the first resolution of the first video data.

**8**. The method of claim **1**, wherein the generated video stream further comprises the first video data, and the generated video stream is configured to allow a remote video conferencing device to simultaneously display the first video data and second video data.

**9**. A system for generating a video stream for use in a video conference, comprising:

a first camera sensor configured to capture first video data comprising a plurality of frames from a first field of view of a physical environment;

a second camera sensor;

a mounting structure capable of adjusting an orientation of the second camera sensor along one or more degrees of freedom;

control logic configured to:

determine activity data for portions of the first video data across the plurality of frames; and

control movement of the mounting structure to adjust the orientation of the second camera along the one or more degrees of freedom, based on the determined activity data; and

video processing logic configured to:

capture second video data from a second field of view of the physical environment using the second camera sensor;

encode the captured second video data;

generate a video stream comprising the captured second video data and audio data captured within the physical environment; and

transmit the video stream to a video conferencing application executing on a user device, wherein the video conferencing application is configured to process the video stream as an input video stream to facilitate the transmission of the video stream to a remote video conferencing device for display.

**10**. The system of claim **9**, wherein the first camera sensor comprises a wide angle camera sensor.

**11**. The system of claim **10**, wherein the control logic configured to determine the activity data for portions of the video data across the plurality of frames is further configured to:

perform a facial detection analysis to detect a plurality of user faces within the first video data; and

determine a measure of motion for each of the detected plurality of user faces across the plurality of frames of the first video data.

**12**. The system of claim **11**, wherein the control logic configured to control movement of the mounting structure to adjust the orientation of the second camera along the one or more degrees of freedom, based on the determined plurality of measures of activity is further configured to:

select one of the plurality of user faces having a corresponding determined measure of motion that is indicative of user speech; and

determine an orientation for capturing video data substantially centered on the selected user face, and

wherein the movement of the mounting structure is controlled such that an orientation of the second camera sensor matches the determined orientation.

**13**. The system of claim **12**, wherein the control logic configured to determine the orientation for capturing the video data substantially centered on the selected user face is further configured to:

identify a physical entity, within the physical environment, corresponding to the selected user face by accessing a model structure describing an attribute of the physical environment; and

determine a direction of the identified physical entity, relative to a physical position of the second camera sensor within the physical environment.

**14**. The system of claim **12**, wherein the system further comprises two or more microphone sensors, and wherein the control logic configured to determine the orientation for capturing the video data substantially centered on the selected user face is further configured to:

collect audio data from the physical environment using the two or more microphone sensors;

identify user speech within the collected audio data; and

determine a direction from which the identified user speech originates, relative to a physical position of the two or more microphone sensors within the physical environment, and

wherein the logic configured to determine the orientation for capturing video data substantially centered on the selected user face operates further based on the determined direction from which the identified user speech originates.

**15**. The system of claim **12**, wherein the generated video stream further comprises the first video data, and the generated video stream is configured to allow a remote video conferencing device to simultaneously display the first video data and second video data.

**16**. A system for generating a video stream for use in a video conference, comprising:

a camera sensor configured to capture first video data comprising a plurality of frames from a first field of view of a physical environment at a first resolution;

control logic configured to:

determine activity data for portions of the first video data across the plurality of frames;

define a portion of the captured first video data to extract, based on the determined activity data; and

extract the portion of the captured video data to create second video data, the second video data having less than all of a plurality of pixels of the captured video data; and

video processing logic configured to:

generate a video stream that comprises the second video data and audio data captured within the physical environment; and

transmit the video stream to a video conferencing application executing on a user device, wherein the video conferencing application is configured to process the video stream as an input video stream to facilitate the transmission of the video stream to a remote video conferencing device for display.

**17**. The system of claim **16**, wherein the control logic configured to determine the activity data for portions of the first video data across the plurality of frames is further configured to:

perform a facial detection analysis to detect a plurality of user faces within the captured first video data; and

determine a measure of motion for each of the detected plurality of user faces across the plurality of frames of the first video data.

**18**. The system of claim **17**, wherein the control logic to determine the portion of the first video data to extract, based on the determined activity data, is further configured to:

select one of the plurality of user faces having a corresponding determined measure of motion that is indicative of user speech; and

determine the portion of the captured video to extract, such that the second video data is substantially centered on the selected user face.

**19**. The system of claim **16**, further comprising two or more microphone sensors, and wherein the control logic configured to determine the portion of the captured first video data to extract, based on the determined activity data, is further configured to:

collecting audio data from the physical environment using the two or more microphone sensors;

identifying user speech within the collected audio data; and

determining a direction from which the identified user speech originates, relative to a physical position of the two or more microphone sensors.

**20**. The system of claim **19**, wherein the control logic configured to determine the portion of the first video data to extract, based on the determined activity data, is further configured to:

identify a physical entity, within the physical environment, located in the determined direction from which the user speech originates, by accessing a mapping structure describing an attribute of the physical environment; and

determine a visual representation of the identified physical entity within the plurality of frames of the first video data.

**21**. The system of claim **16**, wherein the video stream further comprises the first video data, and the transmitted video stream is configured to allow the remote video conferencing device to simultaneously display the first video data and second video data.

**22**. The method of claim **1**, wherein the received first video data is captured at a first resolution, and wherein generating the second video data from the second field of view of the physical environment further comprises:

extracting a portion of the first video data to create the second video data, wherein the second video data has less than all of a plurality of pixels of the first video data.

**23**. A non-transitory computer-readable medium containing computer program code that, when executed by operation of one or more computer processors, performs an operation for generating a video stream for use in a video conference, the operation comprising:

receiving, by a first video conferencing endpoint device, first video data captured from a first field of view of a physical environment, the first video data comprising a plurality of frames;

determining activity data from portions of the first video data using information provided in the plurality of frames;

generating, by the first video conferencing endpoint device, second video data from a second field of view of the physical environment, based on the determined activity data;

generating a video stream that comprises the generated second video data and audio data captured within the physical environment; and

transmitting the video stream to a video conferencing application executing on a user device, wherein the video conferencing application is configured to process the video stream as an input video stream to facilitate the transmission of the video stream to a remote video conferencing device for display.

**24**. The non-transitory computer-readable medium of claim **23**, wherein generating the second video data from the second field of view of the physical environment further comprises controlling movement of a controlled camera device to capture the second video data, and wherein the received first video data is captured using a wide angle camera device, which is distinct from the controlled camera device.

**25**. The non-transitory computer-readable medium of claim **24**, wherein determining the activity data for portions of the first video data using information provided in the plurality of frames further comprises:

performing a facial detection analysis to detect a plurality of user faces within the first video data; and

determining a measure of motion for each of the detected plurality of user faces using information provided in the plurality of frames of the first video data.

**26**. The non-transitory computer-readable medium of claim **25**, wherein generating the second video data from the second field of view of the physical environment further comprises:

selecting one of the plurality of user faces having a corresponding determined measure of motion that is indicative of user speech; and

determining an orientation of the camera device for capturing video data substantially centered on the selected user face, and

wherein controlling the movement of the camera device to capture the second video data further comprises controlling the movement of the camera device to match the determined orientation.

**27**. The non-transitory computer-readable medium of claim **26**, wherein determining an orientation of the camera device for capturing the video data substantially centered on the selected user face further comprises:

identifying a physical entity corresponding to the selected user face by accessing a model structure describing an attribute of the physical environment; and

determining a direction of the identified physical entity, relative to a physical position of the camera device within the physical environment.

**28**. The non-transitory computer-readable medium of claim **26**, wherein generating the second video data from the second field of view of the physical environment further comprises:

collecting audio data from the physical environment using two or more microphone sensors;

identifying user speech within the collected audio data; and

determining a direction from which the identified user speech originates, relative to a physical position of the two or more microphone sensors, and

wherein determining the orientation of the camera device for capturing the video data substantially centered on the selected user face is further based on the determined direction from which the identified user speech originates.

**29**. The non-transitory computer-readable medium of claim **23**, wherein the received first video data is captured at a first resolution, and wherein generating the second video data from the second field of view of the physical environment further comprises:

extracting a portion of the first video data to create the second video data, wherein the second video data has a second resolution that is less than the first resolution of the first video data.

* * * * *